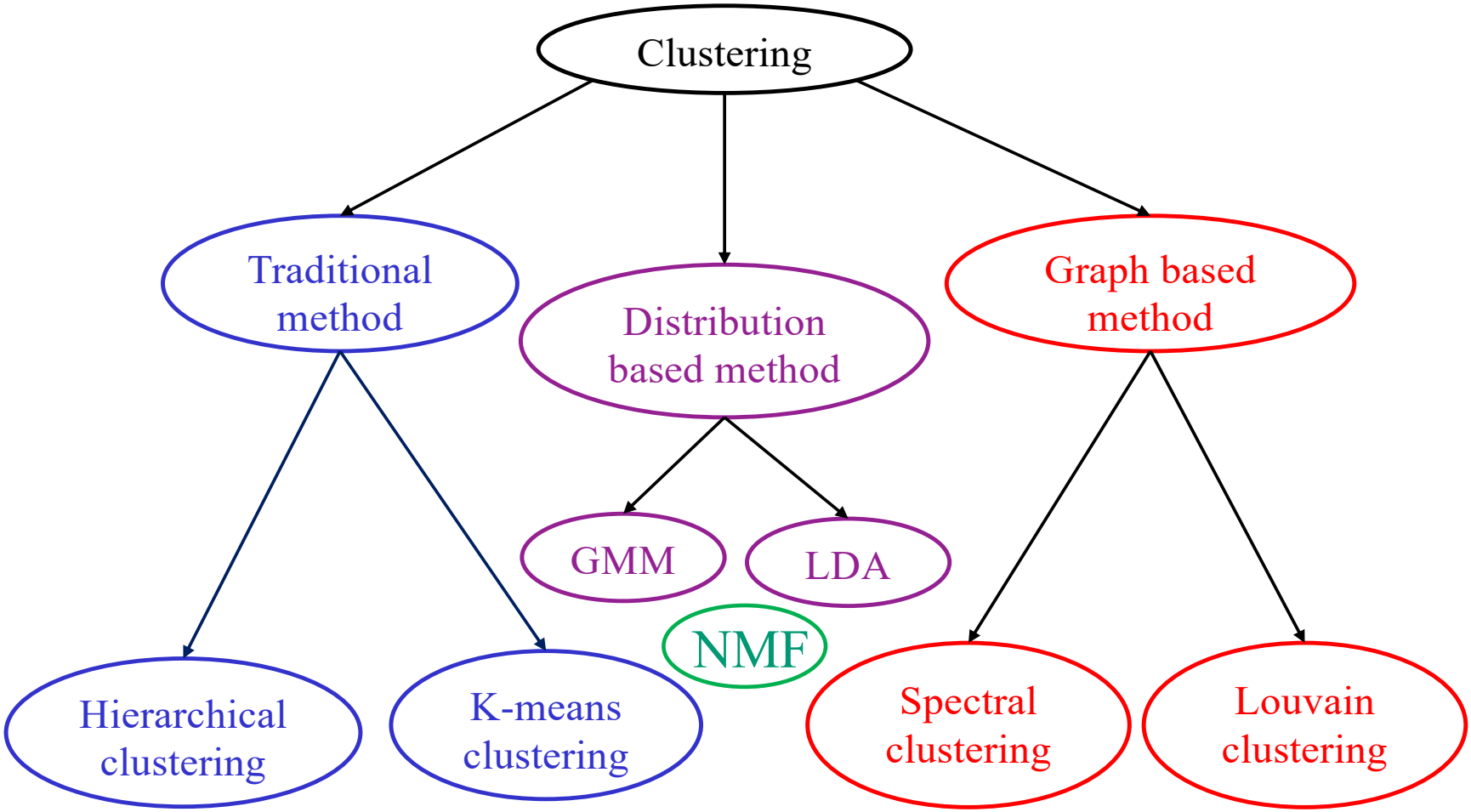
# Clustering Methods:
# From k-means to Gaussian Mixture Model and Louvain Algorithm

Maxwell Lee

High-dimension Data Analysis Group
Laboratory of Cancer Biology and Genetics
Center for Cancer Research
National Cancer Institute

December 7, 2020

# Outline of Clustering Methods

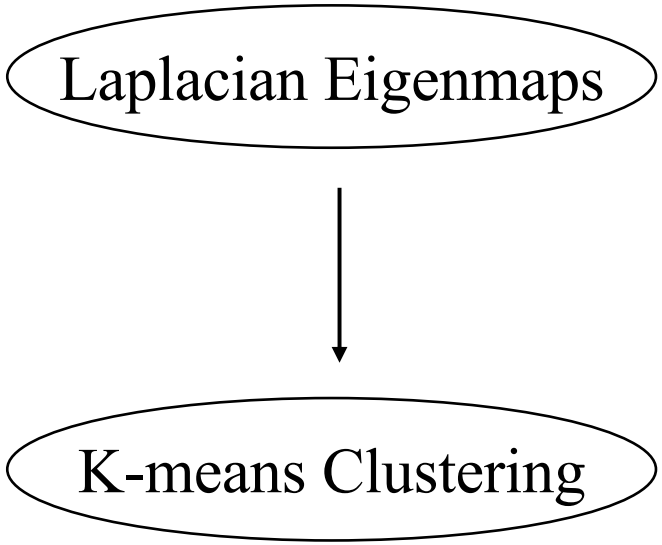



GMM: Gaussian Mixture Model
LDA: Latent Dirichlet Allocation
NMF: Non-negative matrix factorization

Contributed by Emily Tai

# Spectral Clustering

Spectrum: set of its eigenvalues

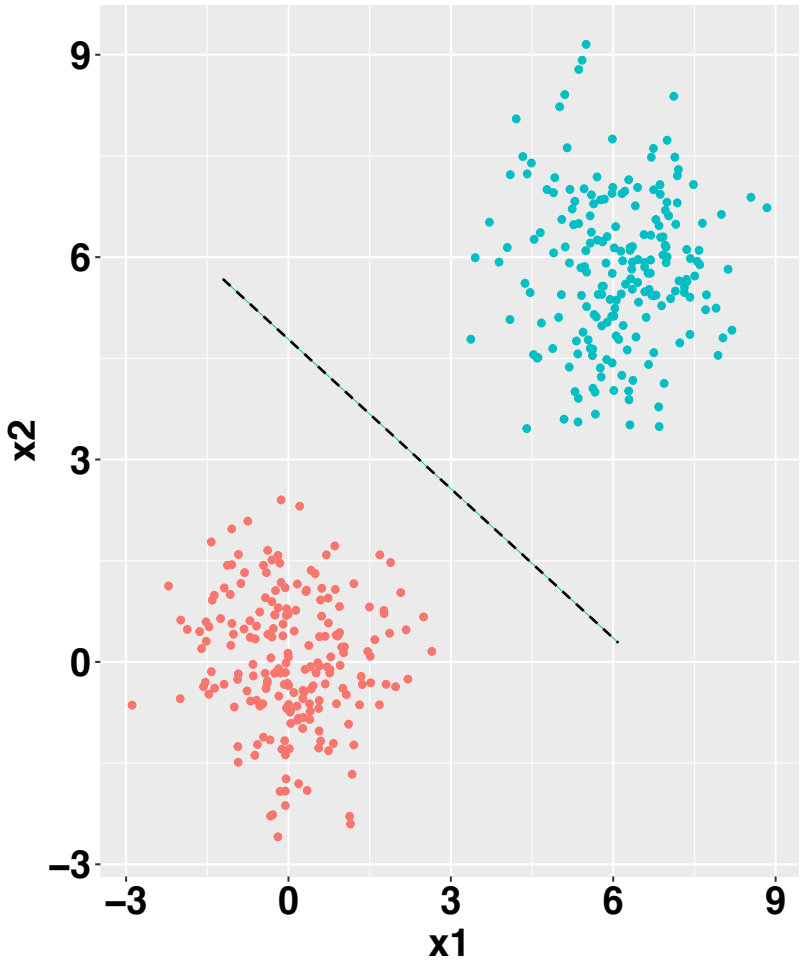



Euclidean distance is not appropriate in HD or in non-Euclidean space

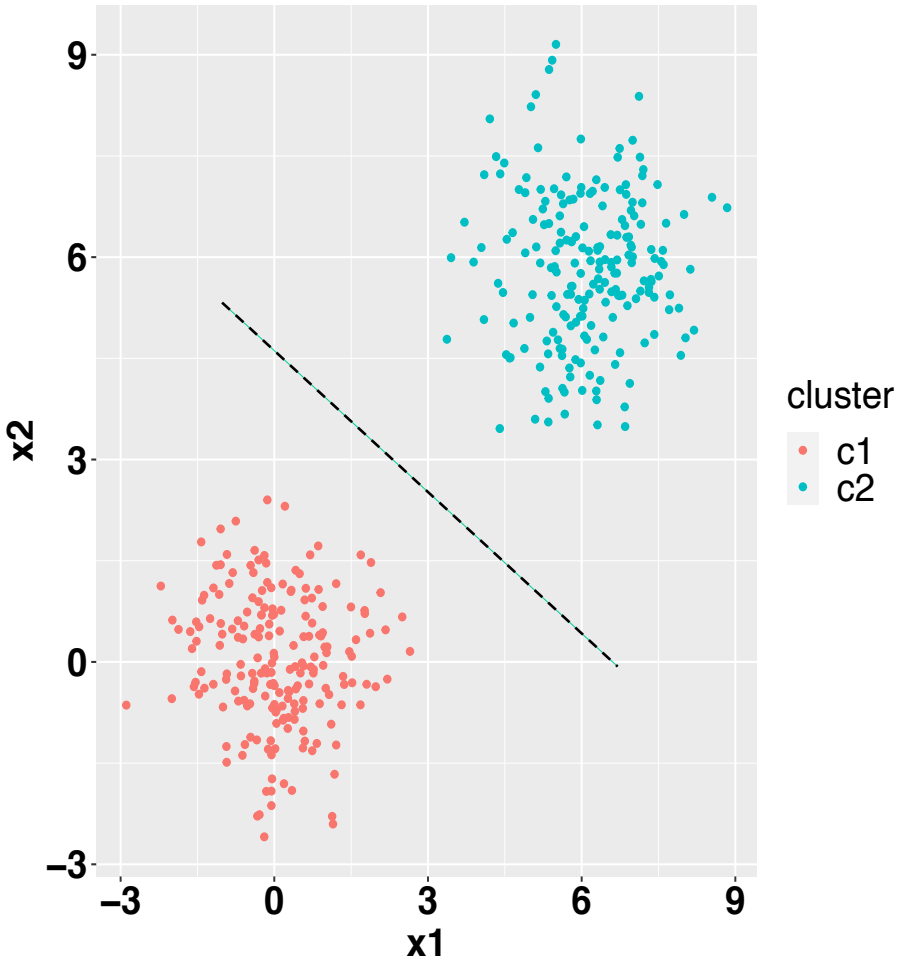
- Curse of dimensionality
- Laplacian Eigenmaps is a non-linear dimension reduction method

# K-means Clustering

color by group

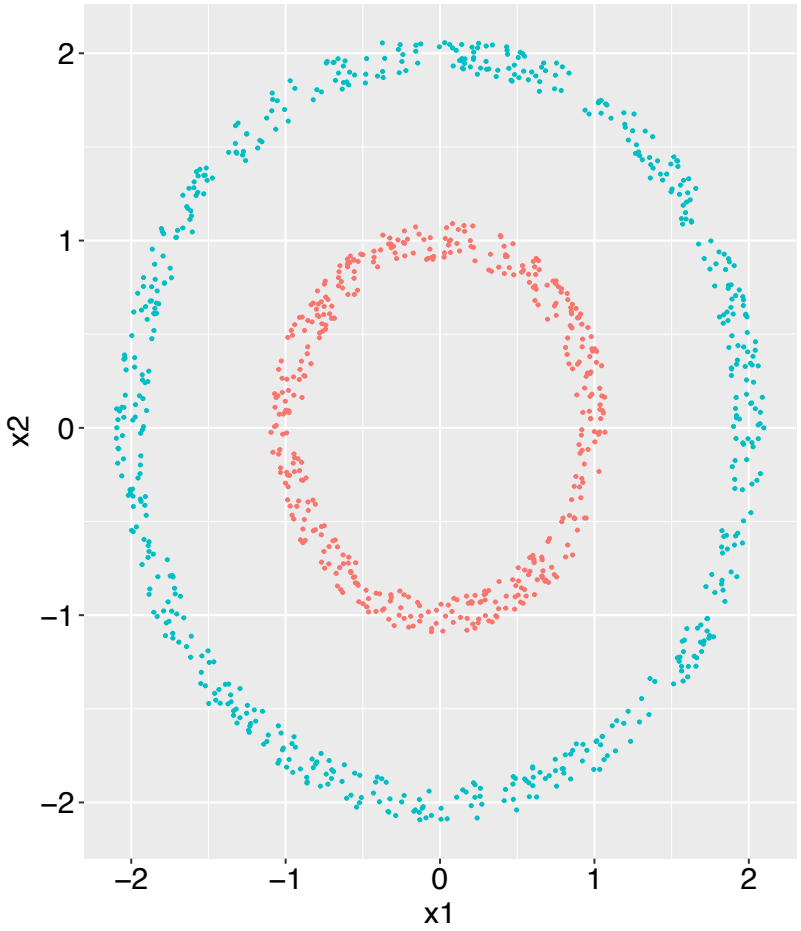



color by k-means cluster

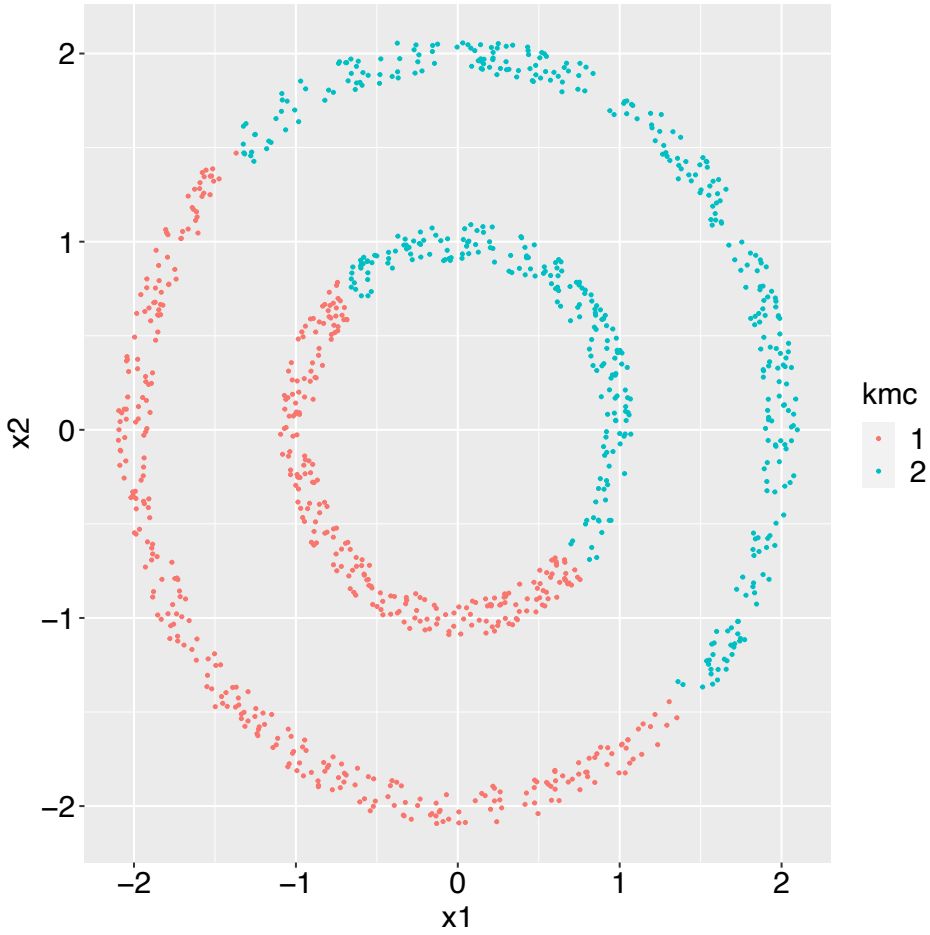


# Two Circles and K-means Clustering

color by group

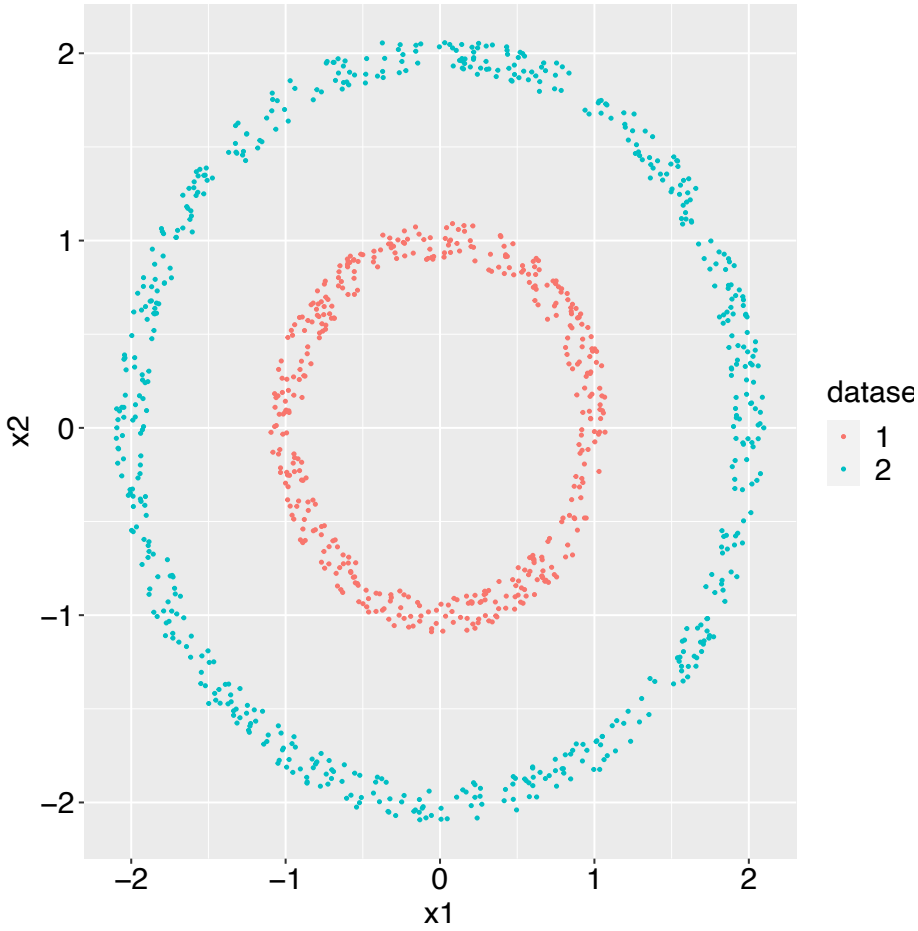



color by k-means cluster

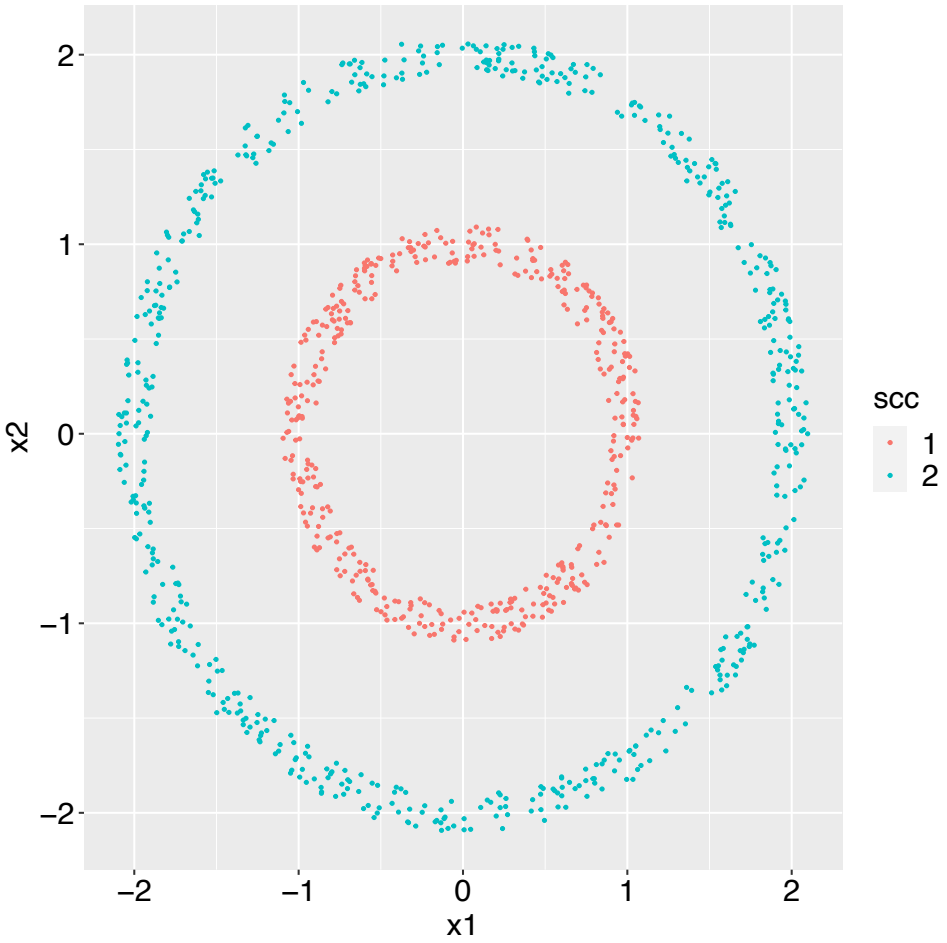


# Two Circles and Spectral Clustering

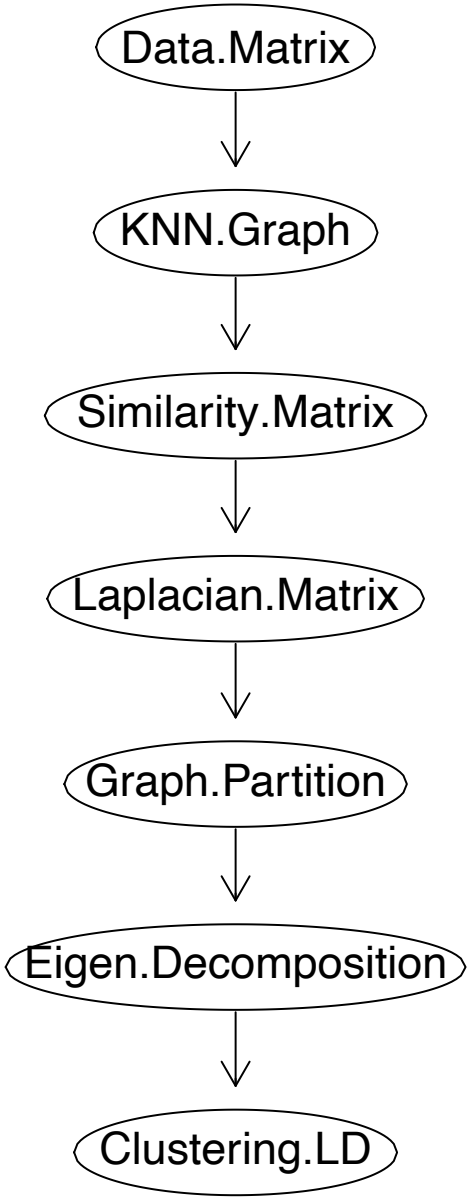
color by group




color by Spectral Clustering



# Algorithm of Spectral Clustering




Laplacian Eigenmaps

KNN: k nearest neighbor

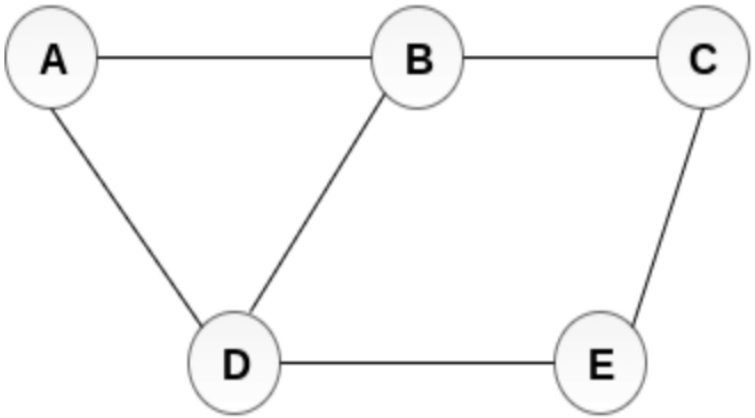
Weight matrix: $w_{ij}=\exp((x_i - x_j)^{\wedge 2}/\sigma)$

diag(W) = (0,0,0,0,0)

L: Laplacian Matrix (D – W)

LD: low dimension

# Adjacency Matrix

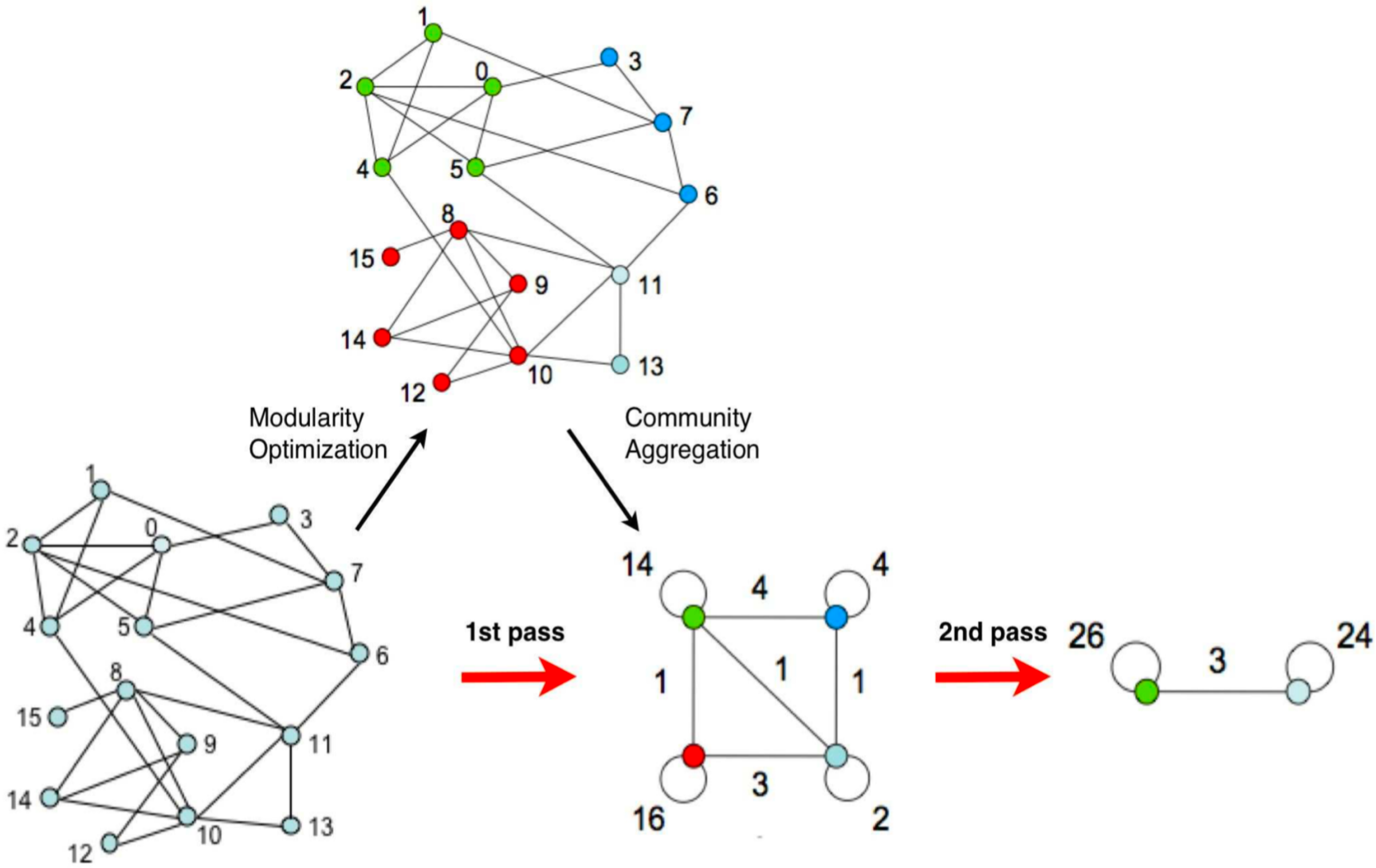



Undirected Graph

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 0 | 1 | 0 | 1 | 0 |
| B | 1 | 0 | 1 | 1 | 0 |
| C | 0 | 1 | 0 | 0 | 1 |
| D | 1 | 1 | 0 | 0 | 1 |
| E | 0 | 0 | 1 | 1 | 0 |

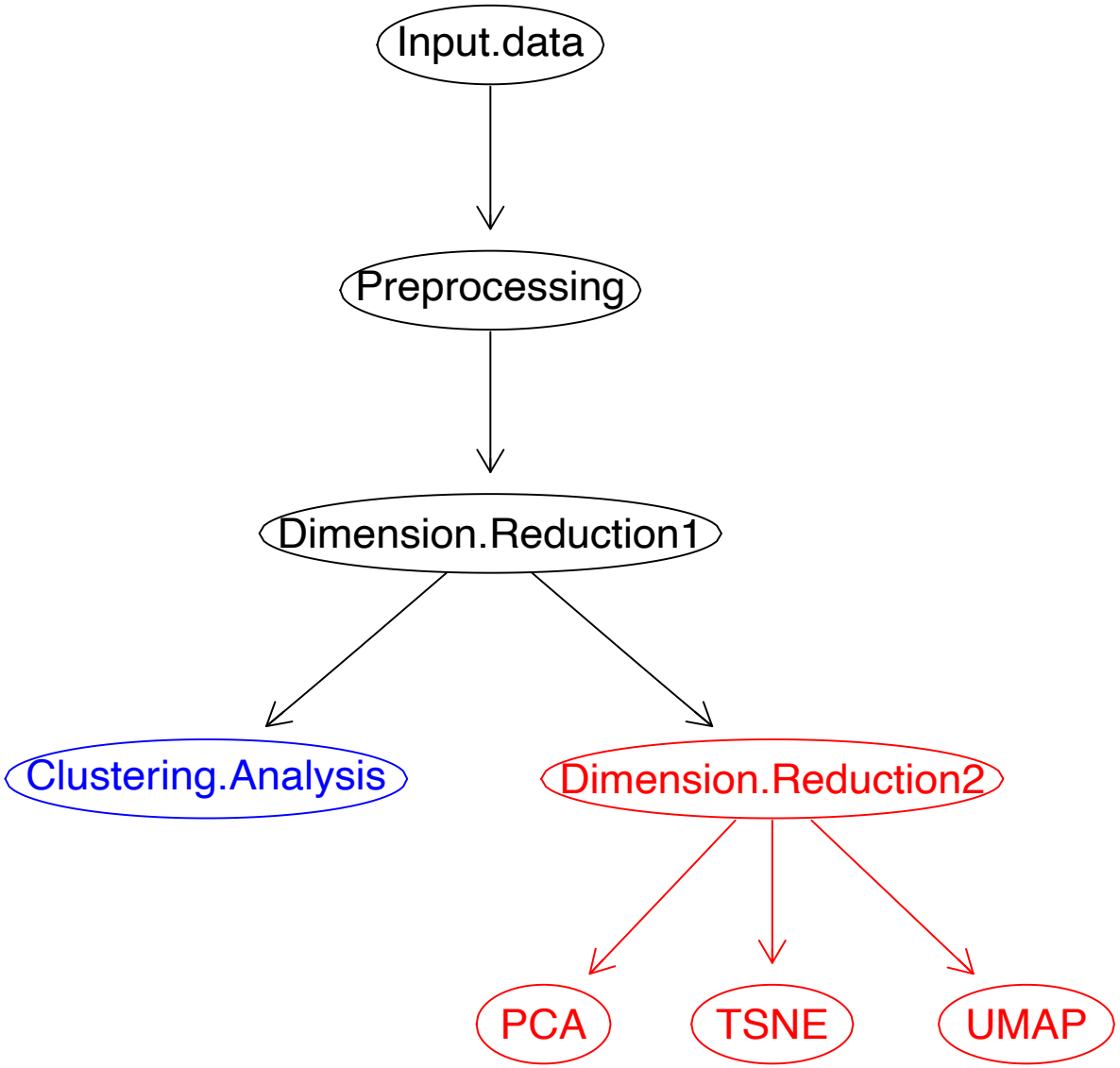
Adjacency Matrix

# Louvain Clustering

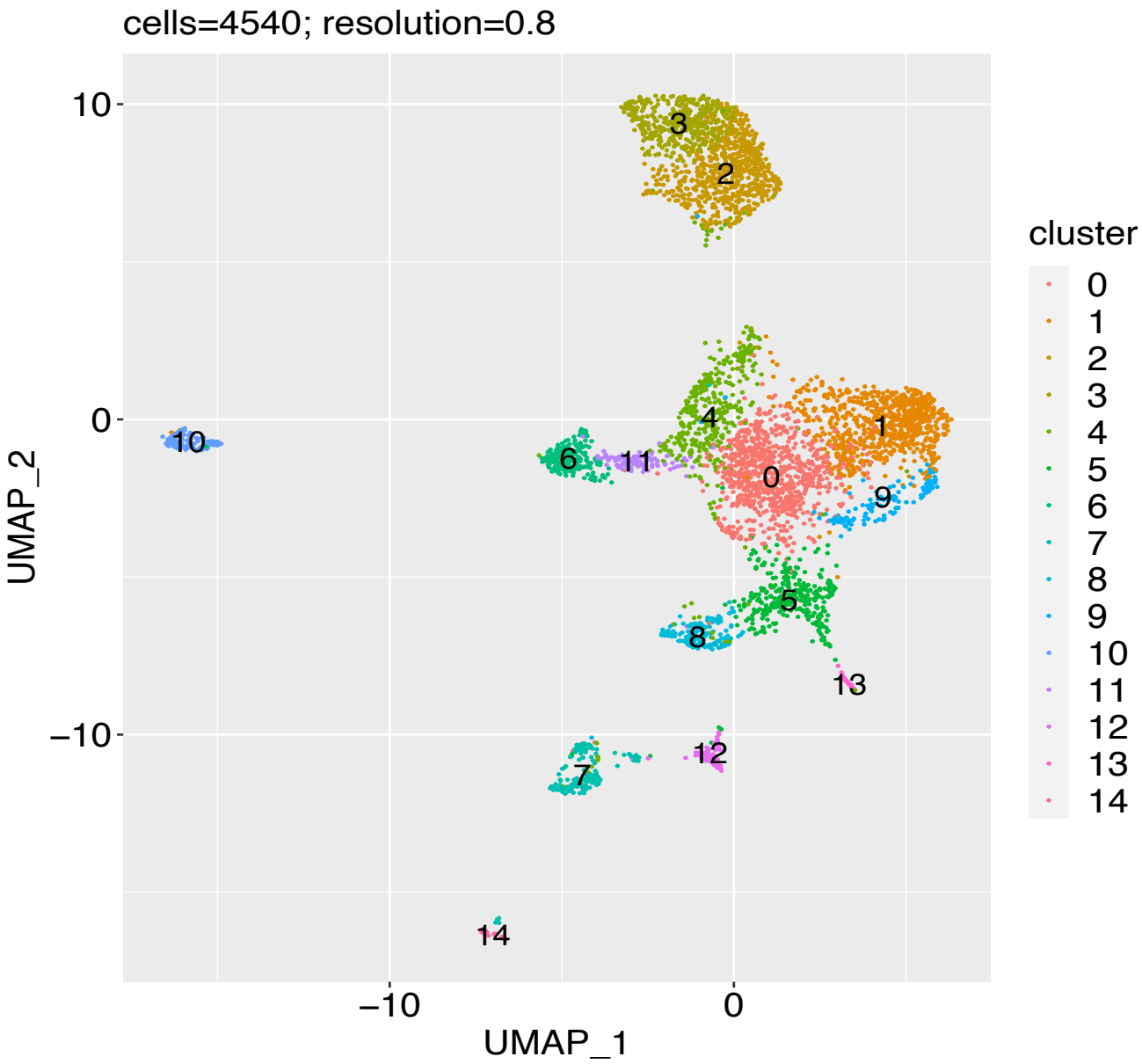



Blondel et al. *J. Stat. Mech. 2008*

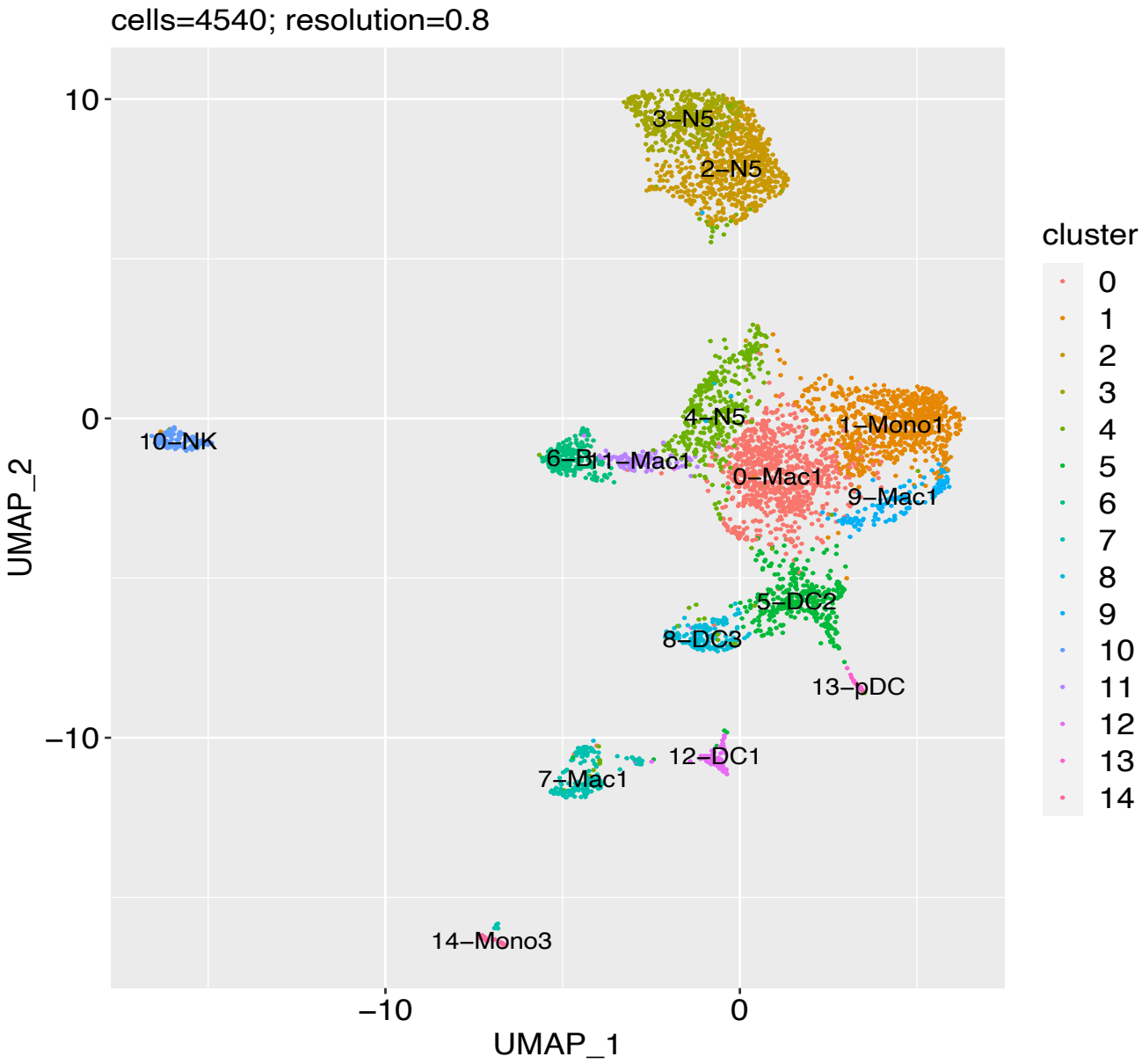
# Flow Chart of scRNAseq Analysis with Seurat Package



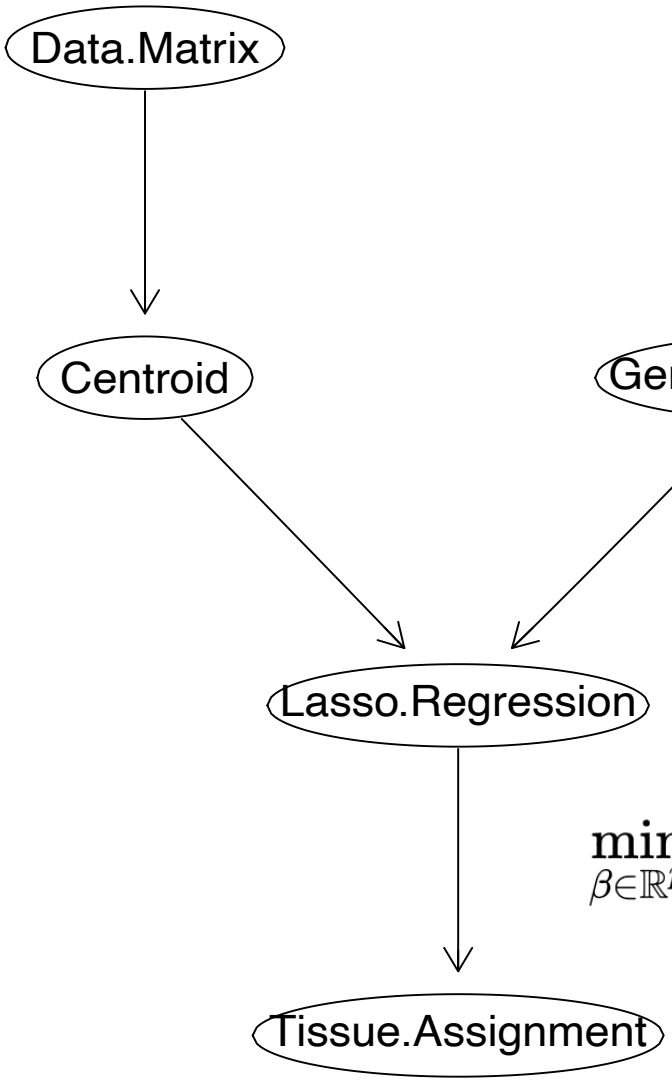
# Louvain Clustering of scRNAseq Data



# Louvain Clustering of scRNAseq Data



# Mapping Cluster to Tissue Type

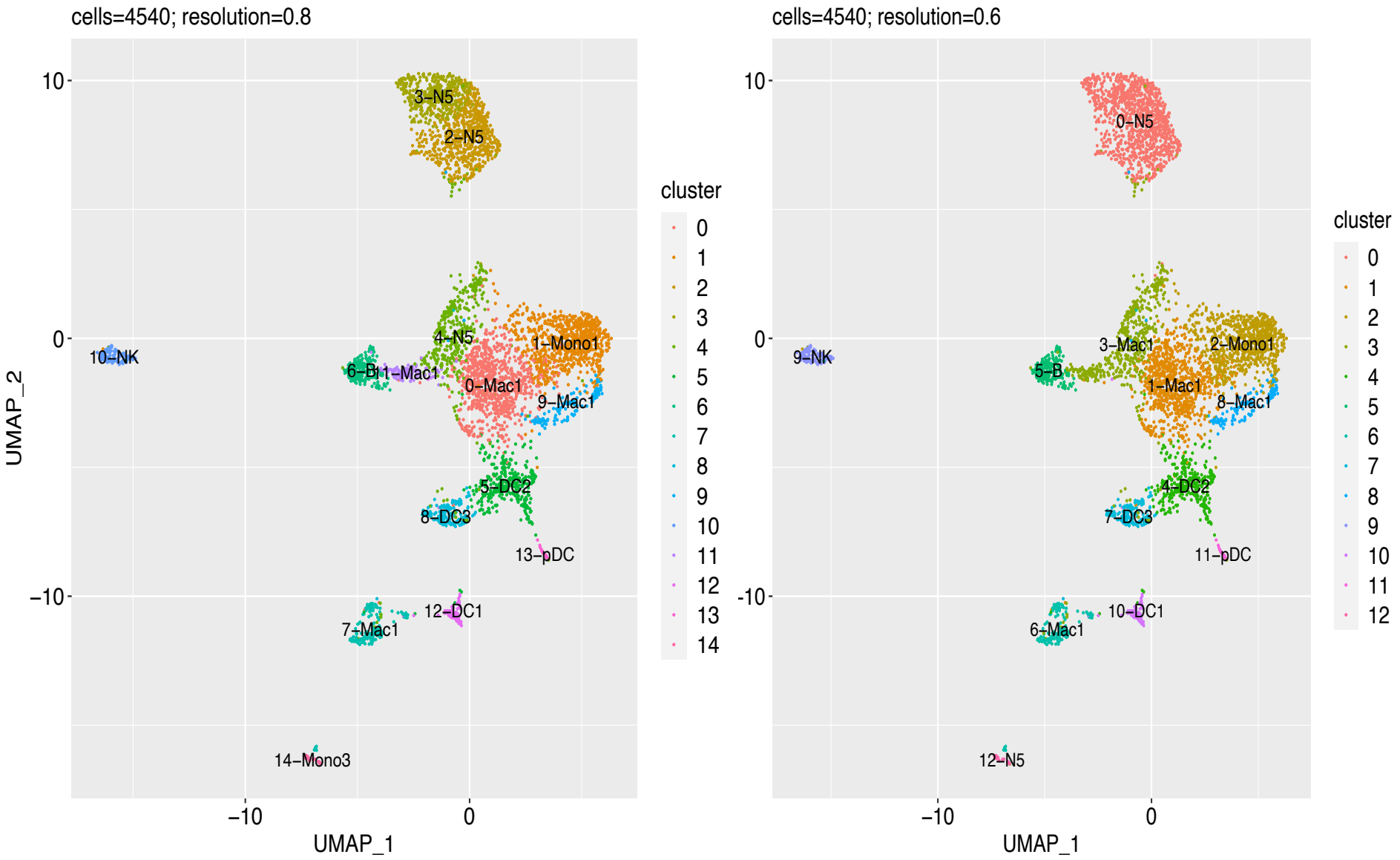



Gene signature:
24 immune cell types
Cibersort LM22 gene list

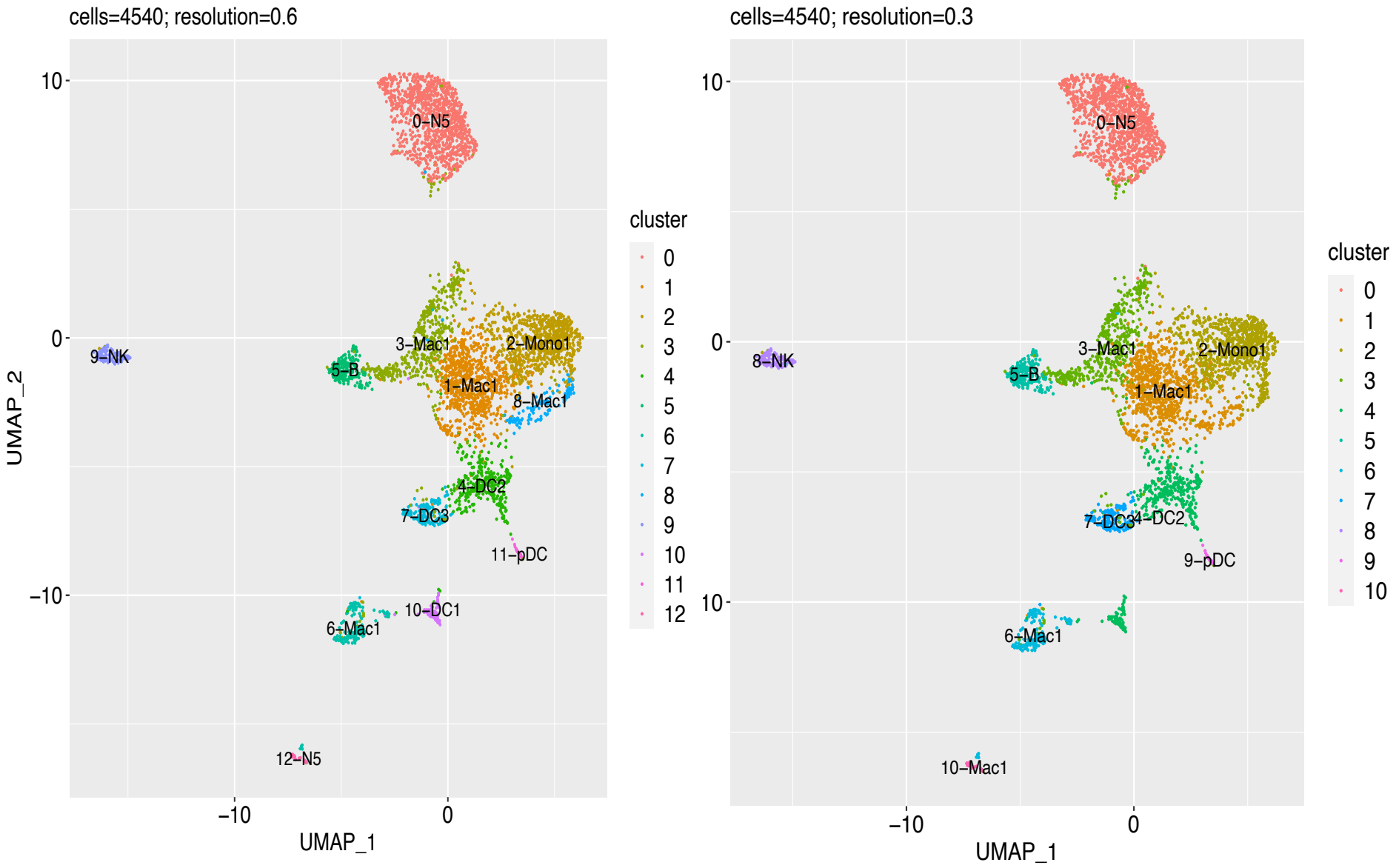
$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

Newman et al Nature Methods 2015; Zilionis et al Immunity 2019

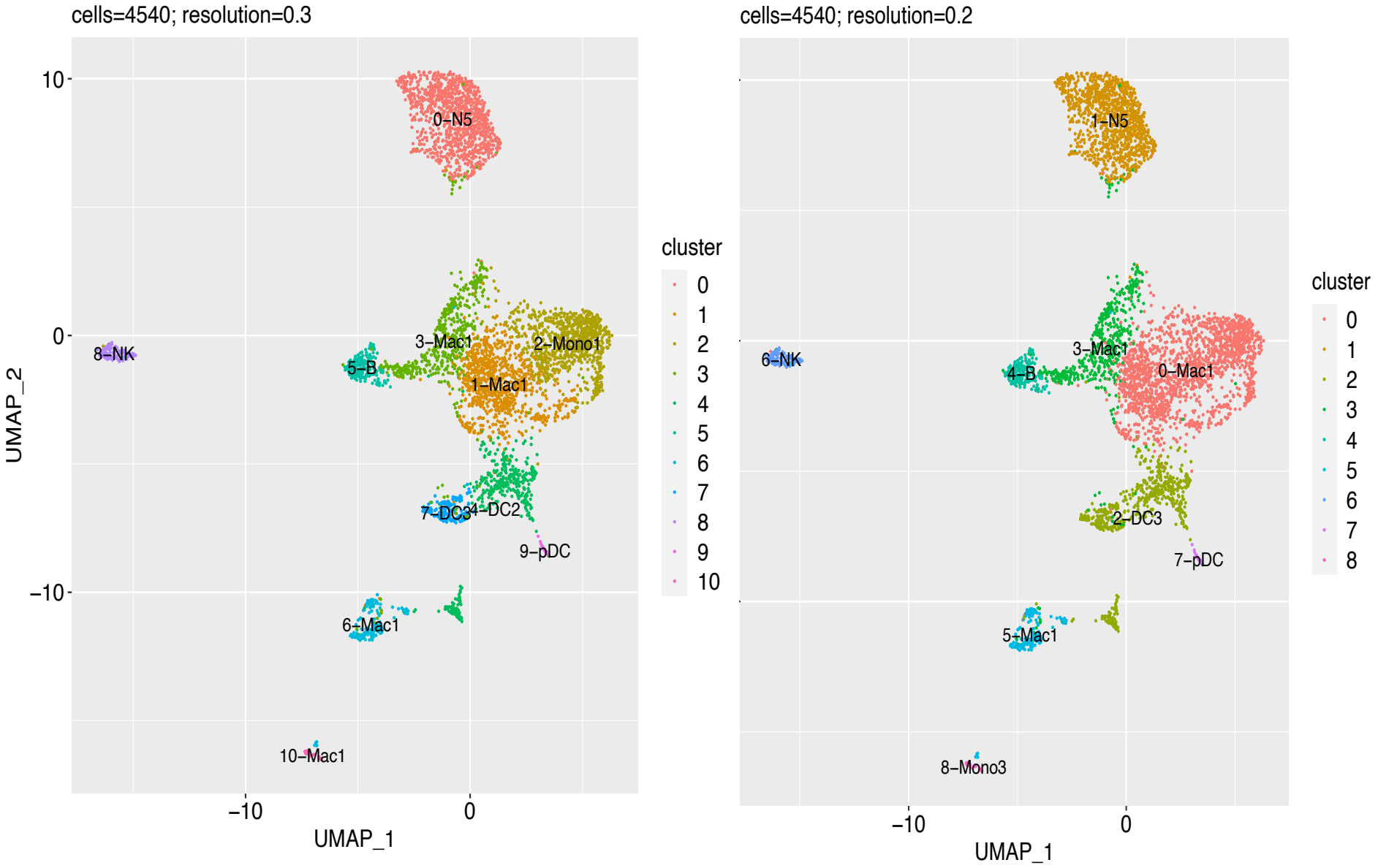
# Louvain Clustering of scRNAseq Data: Resolution



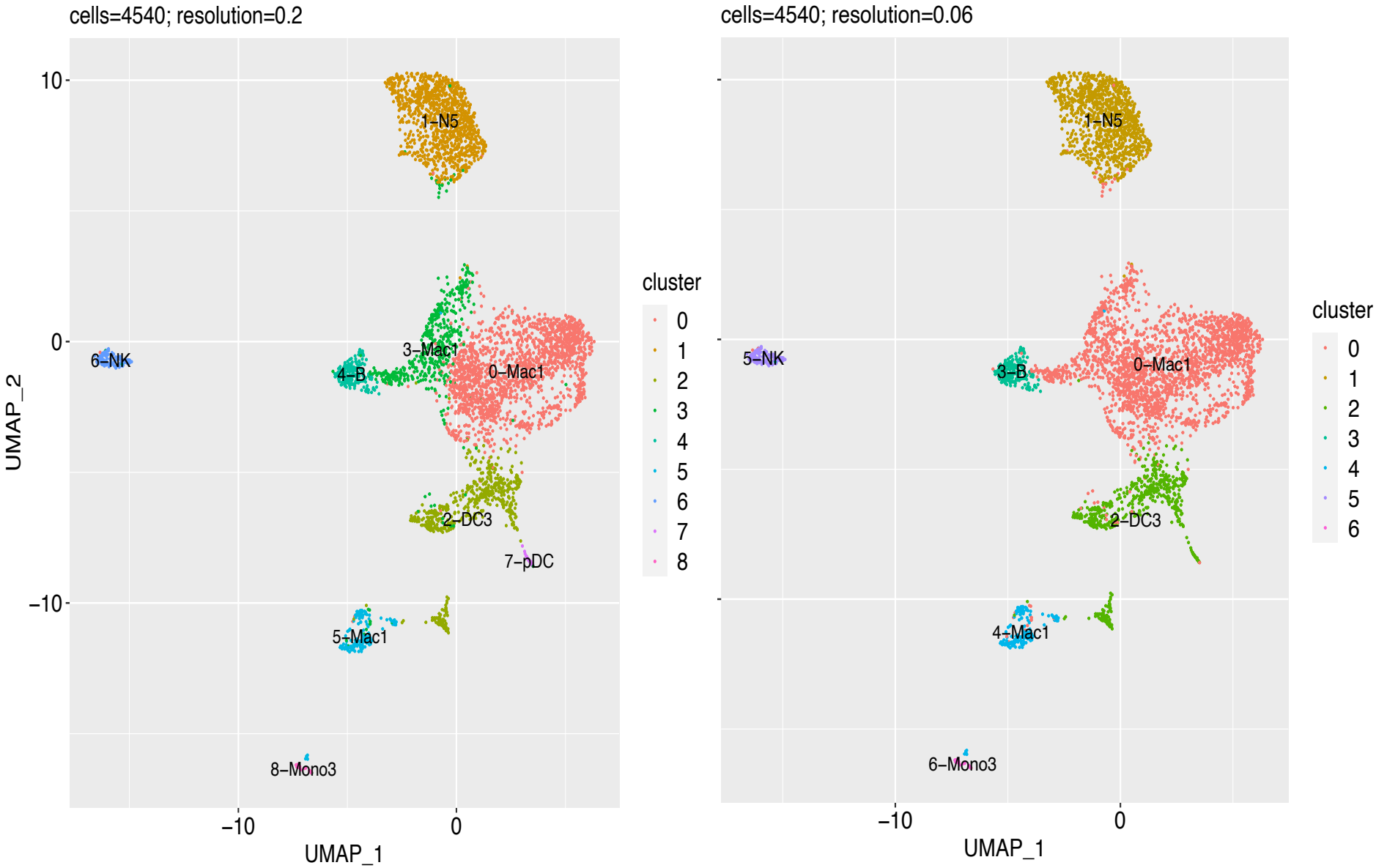
# Louvain Clustering of scRNAseq Data: Resolution



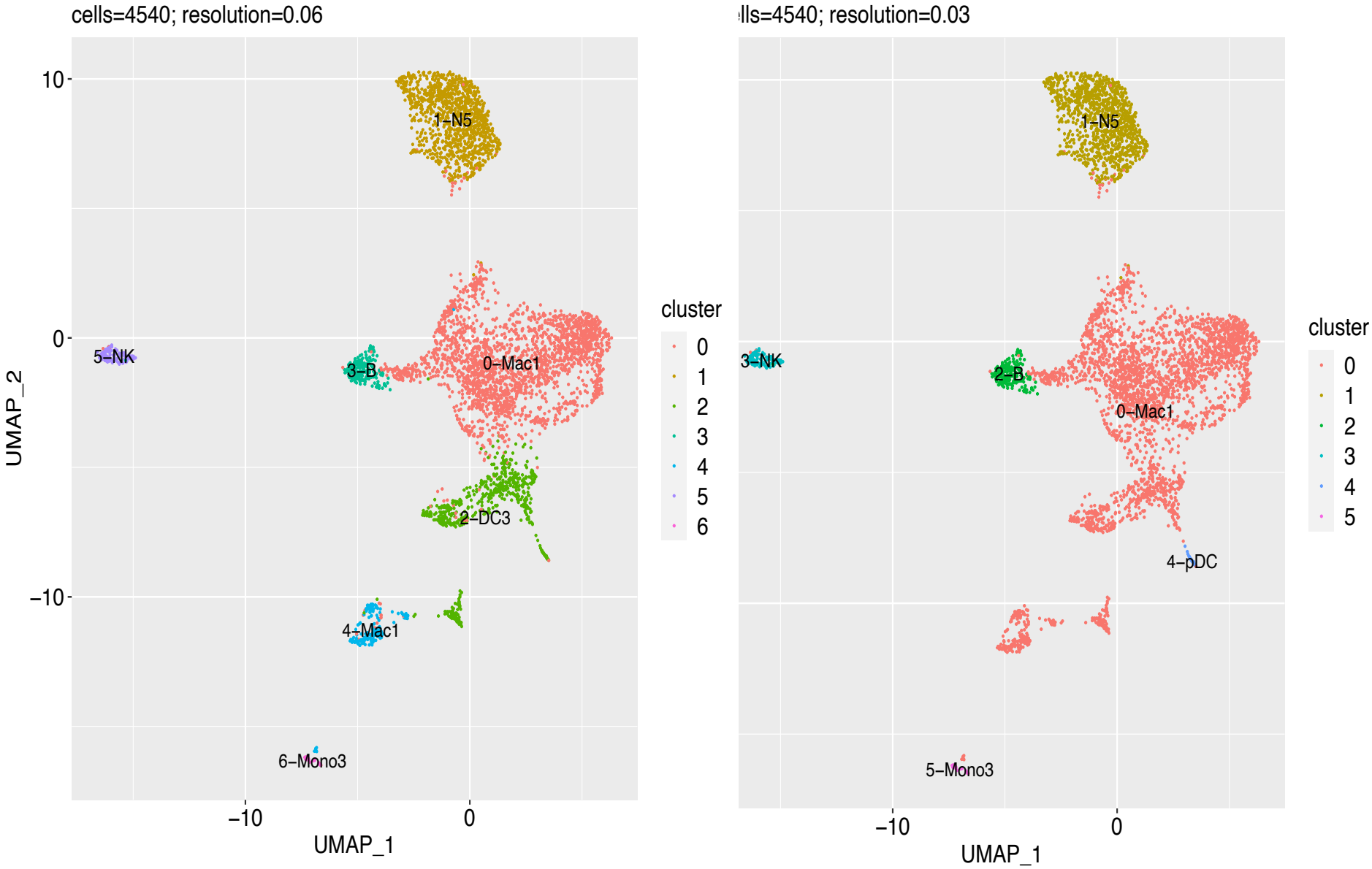
# Louvain Clustering of scRNAseq Data: Resolution



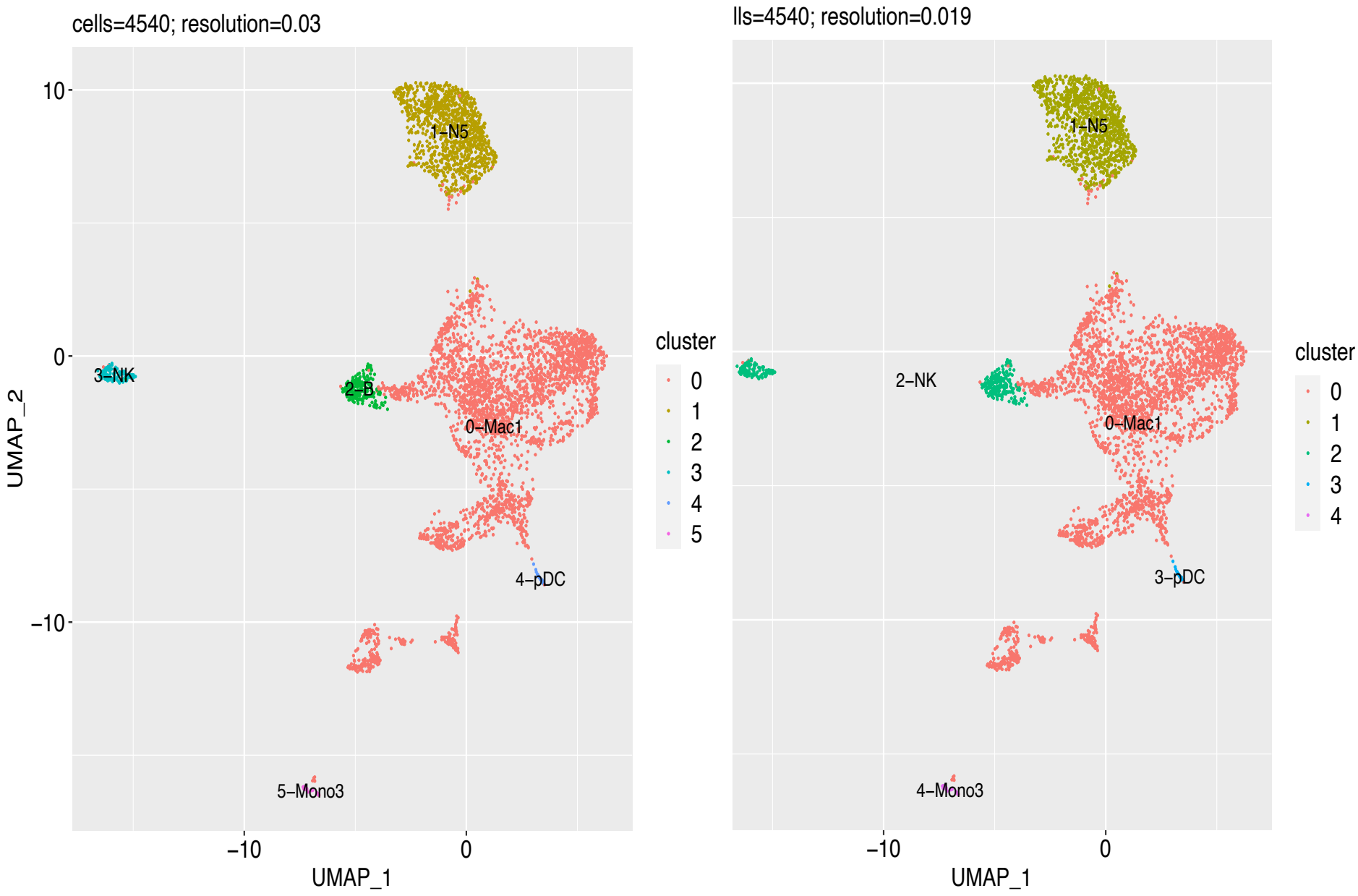
# Louvain Clustering of scRNAseq Data: Resolution



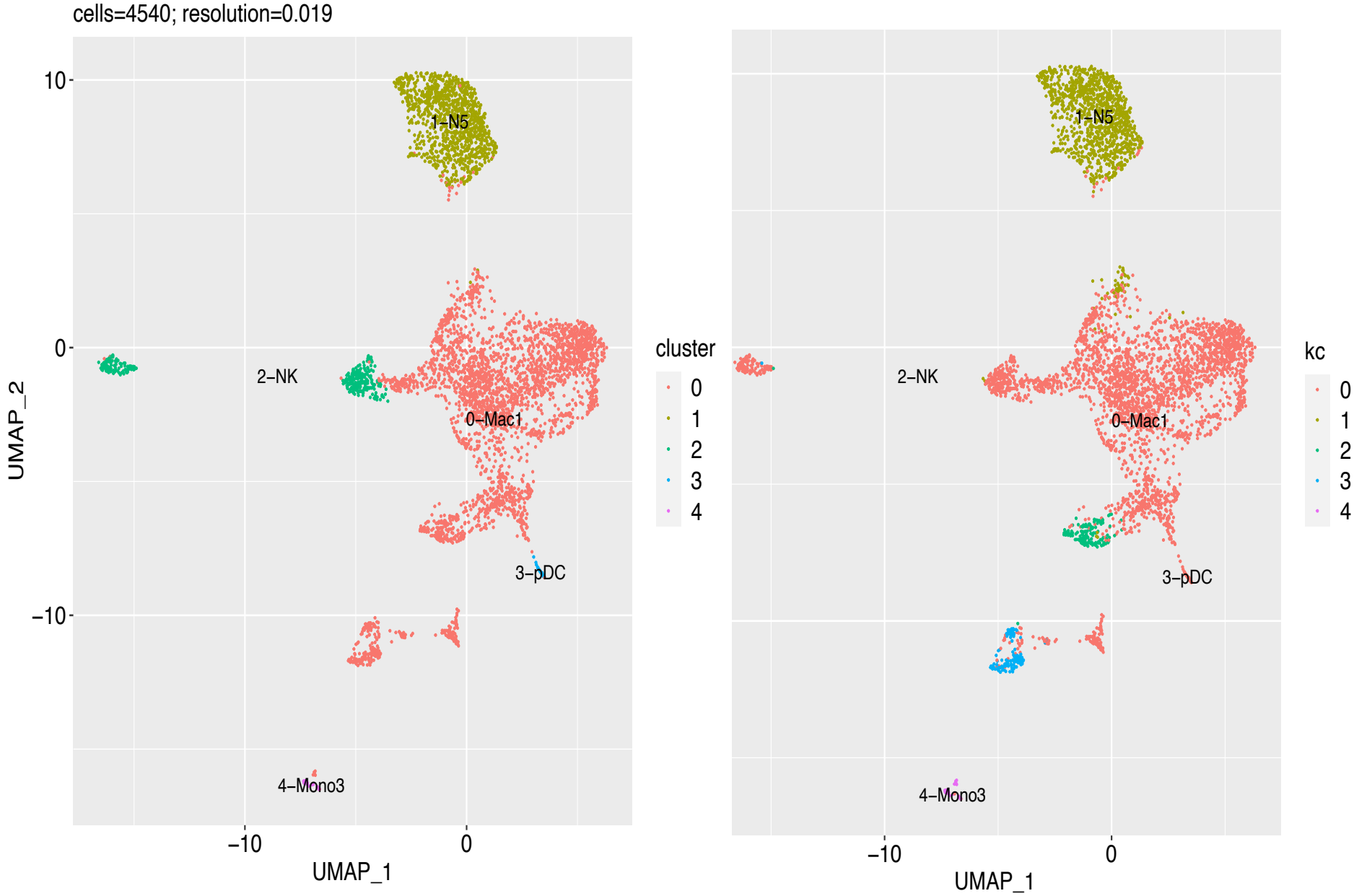
# Louvain Clustering of scRNAseq Data: Resolution



# Louvain Clustering of scRNAseq Data: Resolution



# Louvain Clustering vs. K-means Clustering



# Louvain Clustering vs. k-means Clustering

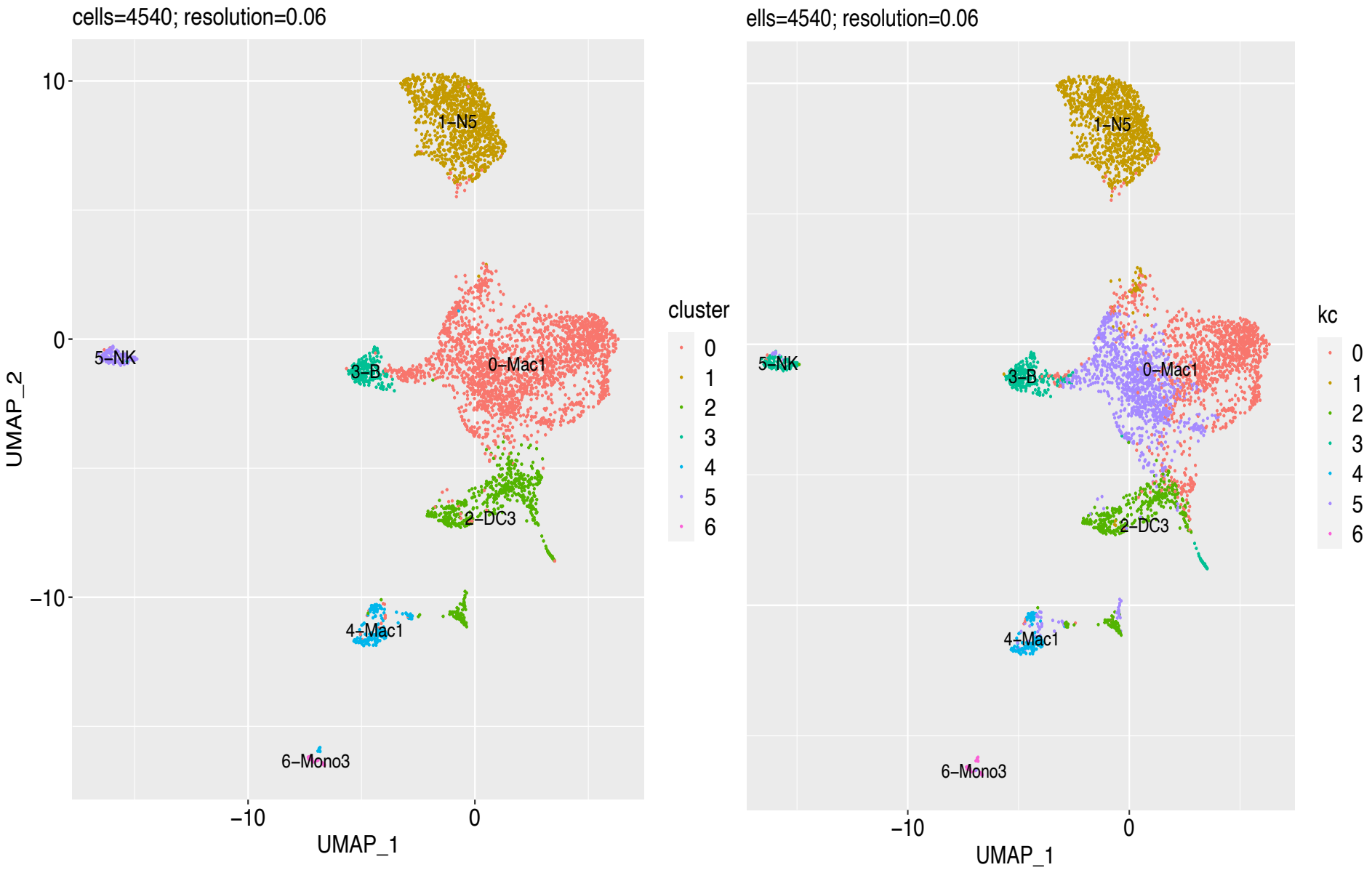
| KC\LC | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | **2694** | 4 | **341** | **36** | 1 |
| 1 | 57 | **1085** | 1 | 0 | 0 |
| 2 | **148** | 0 | 1 | 0 | 0 |
| 3 | **136** | 0 | 1 | 0 | 0 |
| 4 | 7 | 0 | 0 | 0 | **28** |

Accuracy = (2694 + 1085 + 1 + 28) / 4540 = 83.9%

KC: k-means cluster
LC: Louvain cluster

Match
Mismatch

# Louvain Clustering vs. K-means Clustering



# Louvain Clustering vs. k-means Clustering

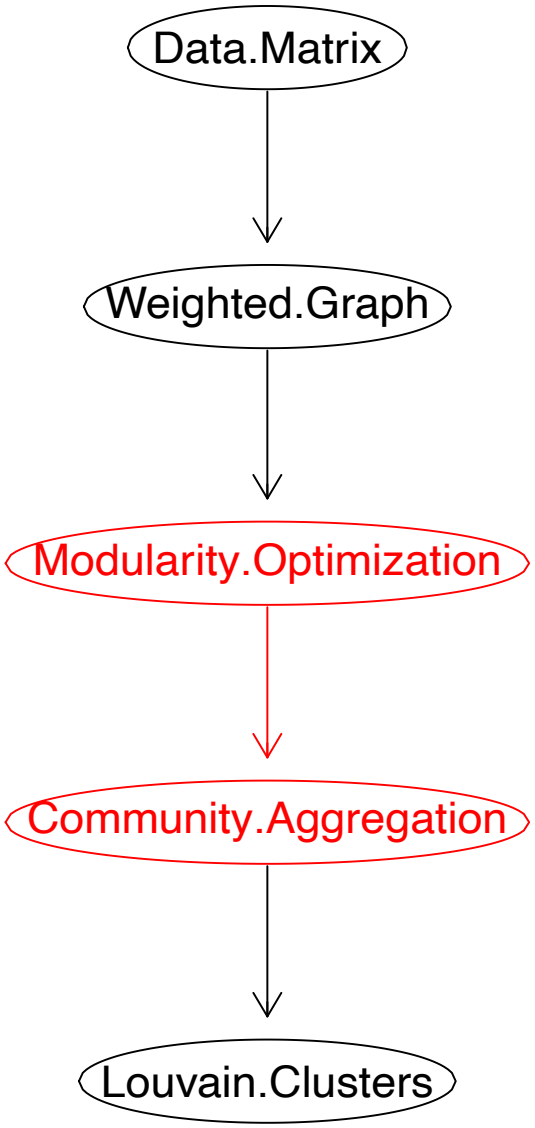
| KC\LC | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 0 | 1201 | 4 | 95 | 0 | 0 | 1 | 0 |
| 1 | 47 | 1085 | 1 | 0 | 0 | 0 | 0 |
| 2 | 3 | 0 | 423 | 0 | 11 | 1 | 0 |
| 3 | 31 | 0 | 37 | 228 | 0 | 107 | 0 |
| 4 | 0 | 0 | 0 | 0 | 133 | 0 | 0 |
| 5 | **1017** | 0 | 46 | 0 | 25 | 7 | 1 |
| 6 | 0 | 0 | 1 | 0 | 7 | 0 | 28 |

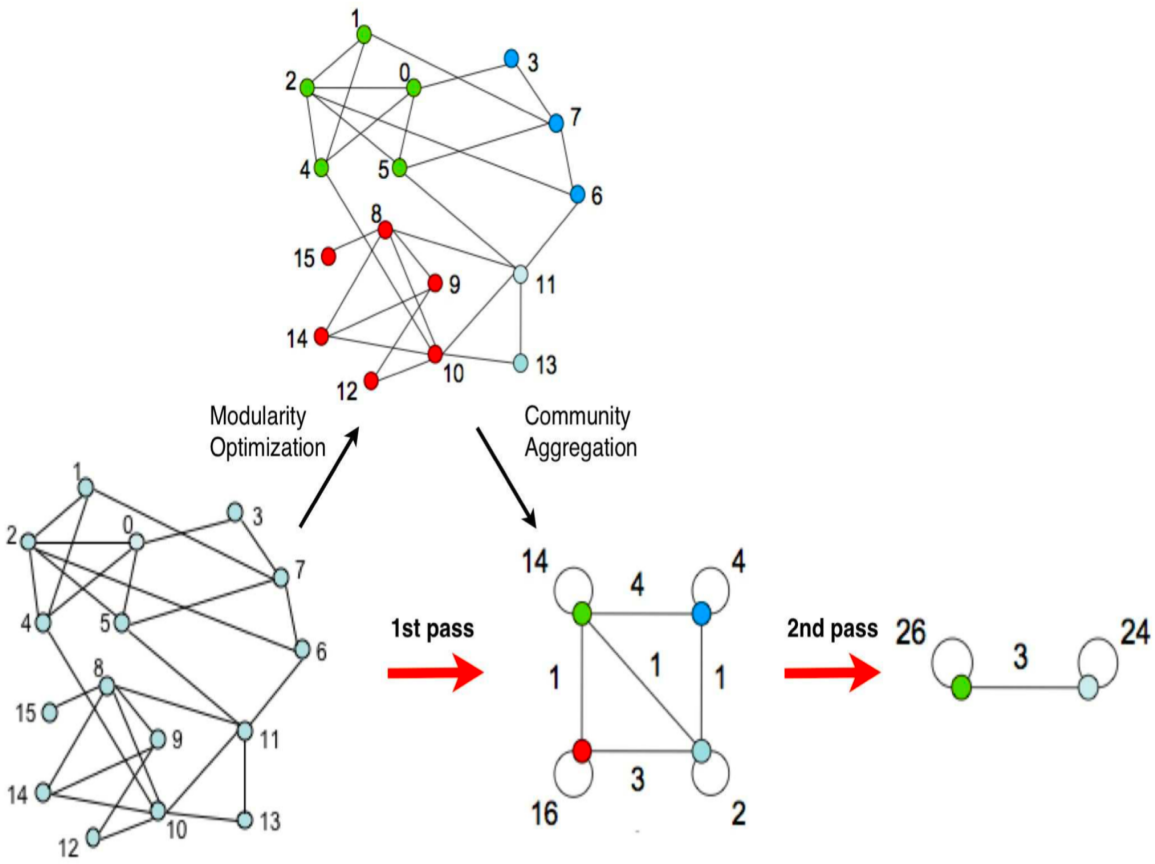
Accuracy = (1201 + 1085 + 423 + 228 + 133 + 7 + 28) / 4540 = 68.4%

KC: k-means cluster
LC: Louvain cluster

Match
Mismatch

# Algorithm of Louvain Clustering

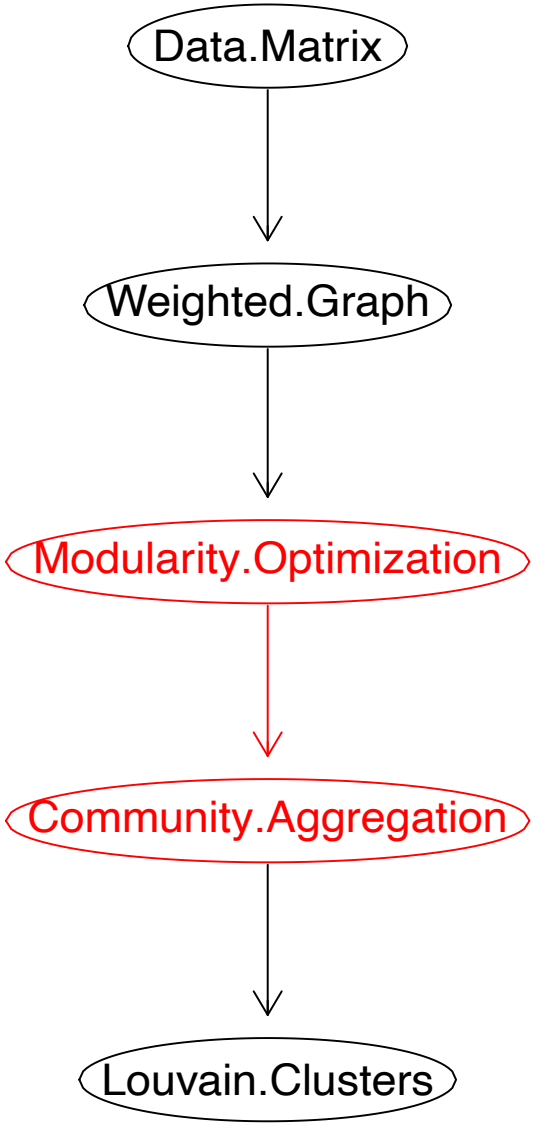


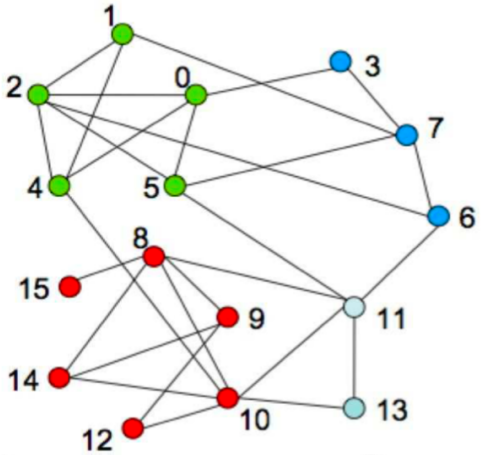




The passes are repeated iteratively until no further increase of modularity

Blondel et al. *J. Stat. Mech. 2008*

# Algorithm of Louvain Clustering




$$Q = \frac{1}{2m} \sum_{i,j} \left[ A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

$A_{ij}$: edge weight between nodes i and j

$k_i$: sum of edge weights attached to node I

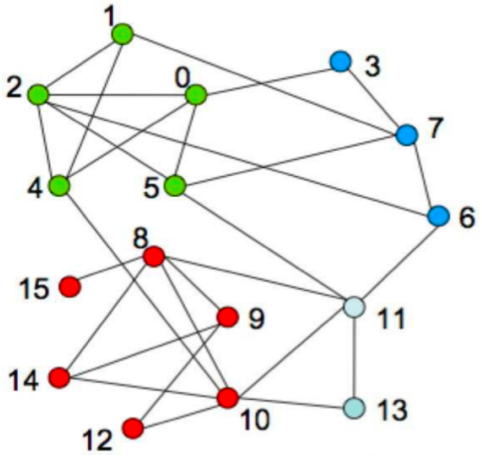
m: sum of all of the edge weights in the graph

$c_i$: communities of the nodes

$\delta(c_i,c_j)=1$ if i and j in the same community, 0 otherwise

Blondel et al. *J. Stat. Mech. 2008*

# Algorithm of Louvain Clustering




$$Q = \frac{1}{2m} \sum_{i,j} \left[ A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

$$\Delta Q = \left[ \frac{\sum_{in} + k_{i,in}}{2m} - \left( \frac{\sum_{tot} + k_i}{2m} \right)^2 \right] - \left[ \frac{\sum_{in}}{2m} - \left( \frac{\sum_{tot}}{2m} \right)^2 - \left( \frac{k_i}{2m} \right)^2 \right]$$

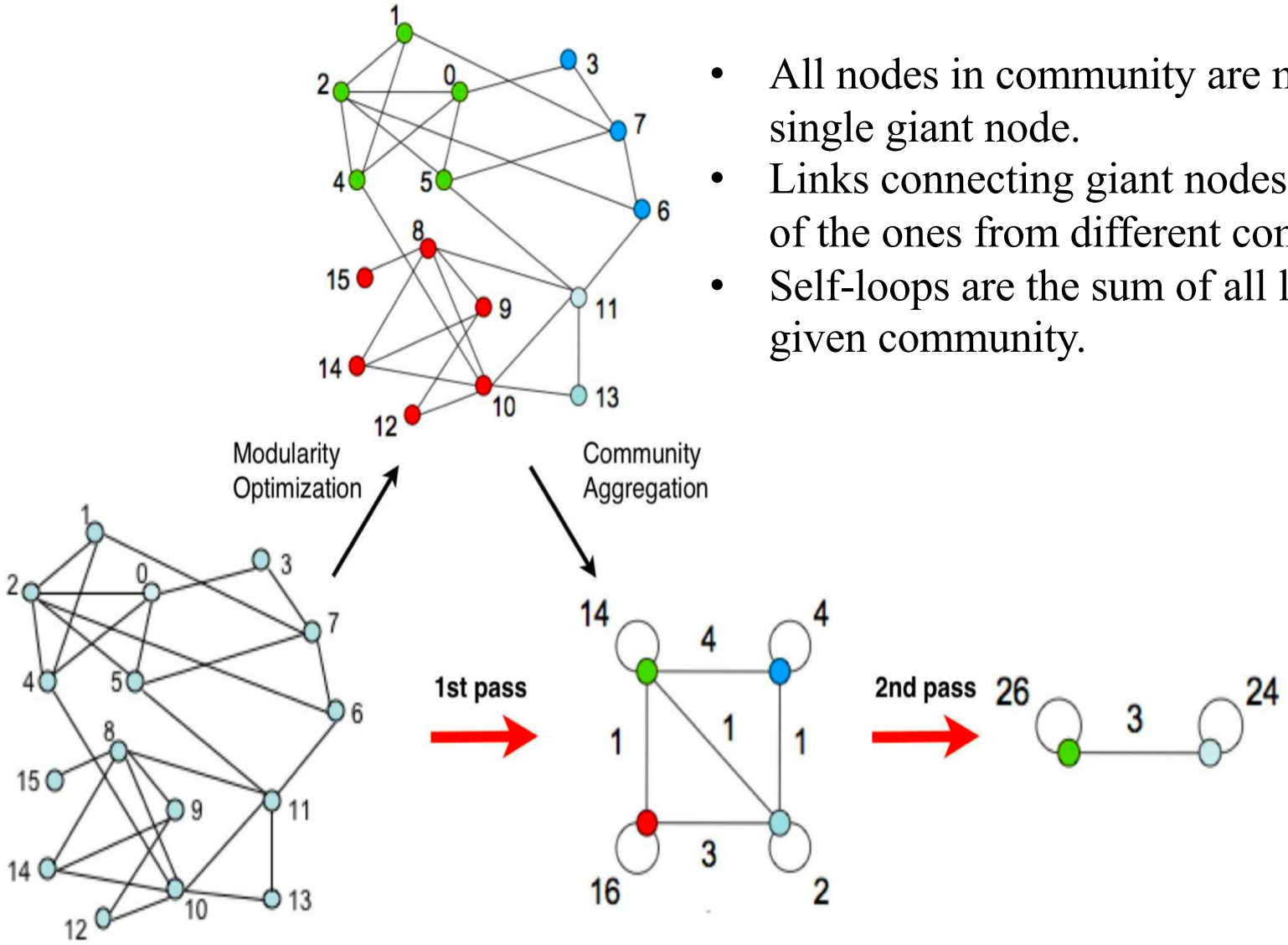
$\Sigma_{in}$: is sum of all the weights of the links inside the community C where the node i is moving into.

$\Sigma_{tot}$: is sum of all the weights of the links incident to the nodes in C

$k_i$: sum of all weights of the links incident to the node i.

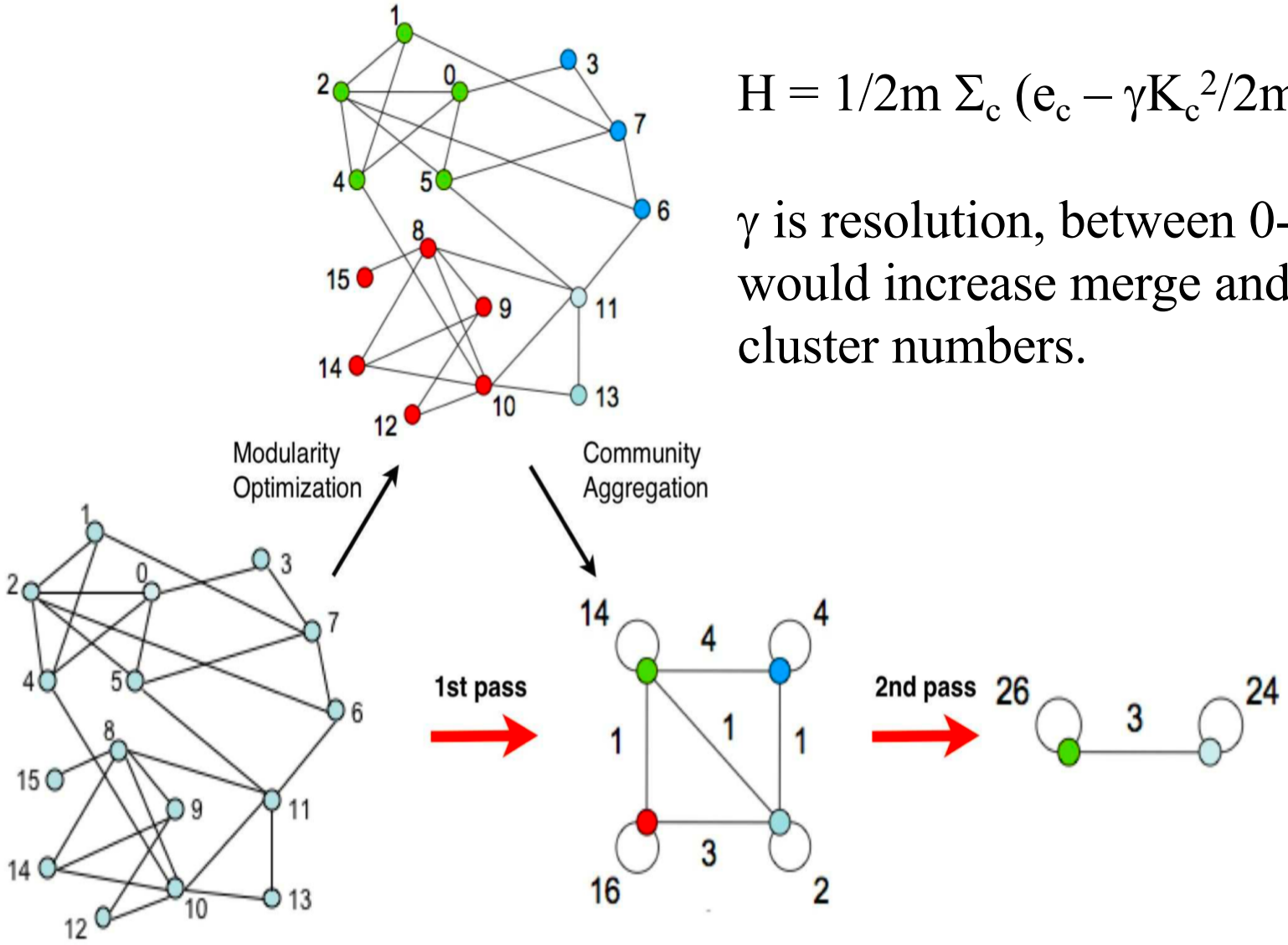
$k_{i,in}$: sum of all weights of links from node I to the nodes in C

# Algorithm of Louvain Clustering




- All nodes in community are merged into a single giant node.
- Links connecting giant nodes are the sum of the ones from different communities.
- Self-loops are the sum of all links inside a given community.

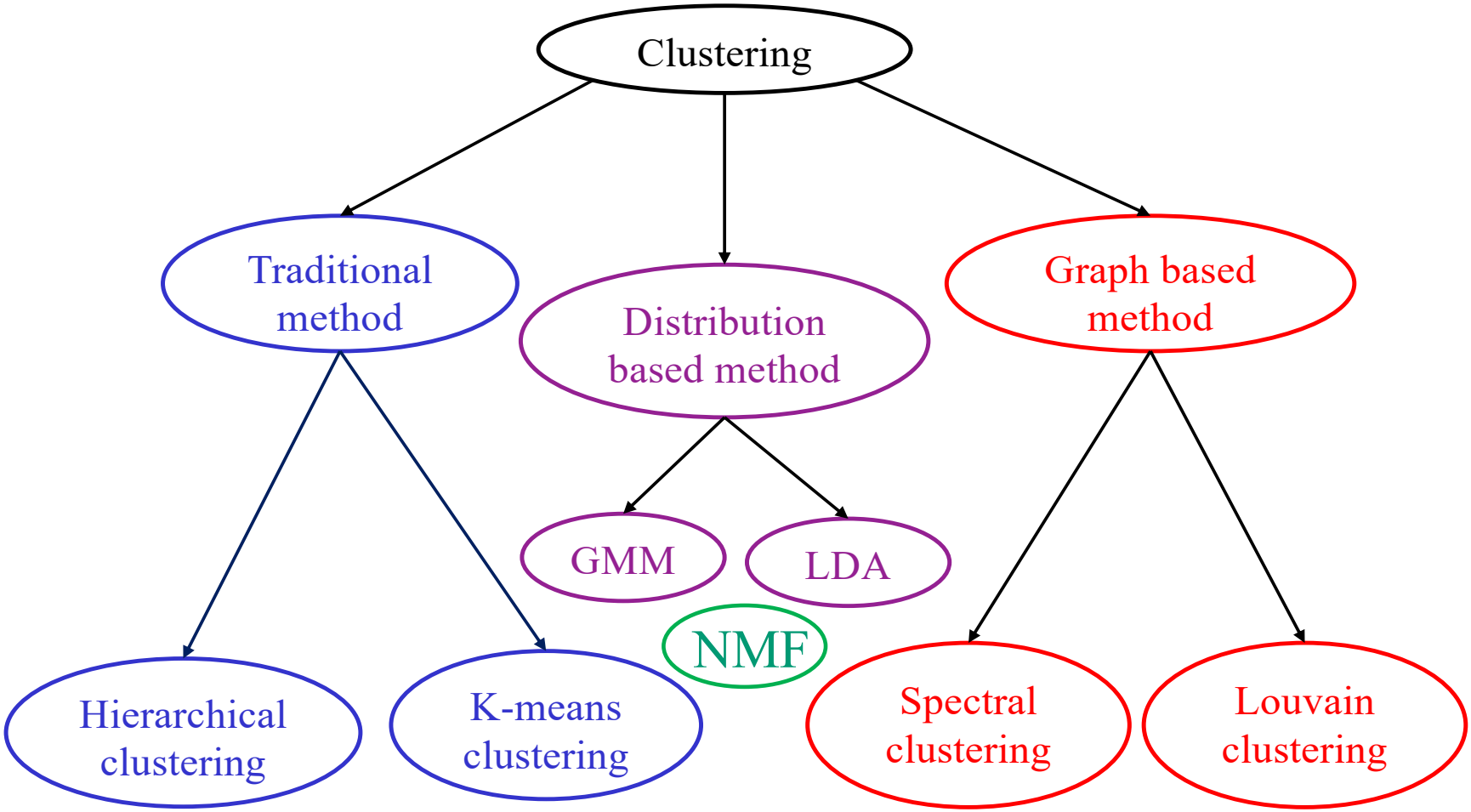
# Algorithm of Louvain Clustering




$$H = 1/2m \, \Sigma_c \, (e_c - \gamma K_c^2/2m)$$

$\gamma$ is resolution, between 0-1, small $\gamma$ would increase merge and reduce cluster numbers.

# Outline of Clustering Methods




GMM: Gaussian Mixture Model
LDA: Latent Dirichlet Allocation
NMF: Non-negative matrix factorization

Contributed by Emily Tai